

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2001年4月26日 (26.04.2001)

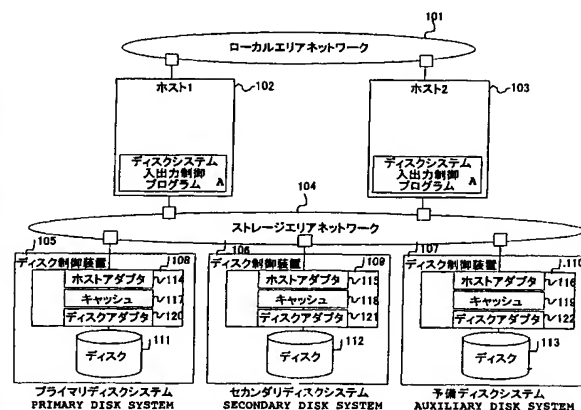
PCT

(10) 国際公開番号
WO 01/29647 A1

- (51) 国際特許分類7: G06F 3/06, 研究所内 Tokyo (JP). 佐藤孝夫 (SATO, Takao) [JP/JP];
G11B 20/10, G06F 12/00, H04L 29/14 千256-8510 神奈川県小田原市国府津2880番地 株
式会社 日立製作所 ストレージシステム事業部内
Kanagawa (JP).
- (21) 国際出願番号: PCT/JP99/05850
- (22) 国際出願日: 1999年10月22日 (22.10.1999) (74) 代理人: 弁理士 作田康夫 (SAKUTA, Yasuo); 千100-
8220 東京都千代田区丸の内一丁目5番1号 株式会社
日立製作所内 Tokyo (JP).
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語 (81) 指定国 (国内): CN, JP, KR, US.
- (71) 出願人 (米国を除く全ての指定国について): 株式会 社 日立製作所 (HITACHI, LTD.) [JP/JP]; 千101-8010
東京都千代田区神田駿河台四丁目6番地 Tokyo (JP). (84) 指定国 (広域): ヨーロッパ特許 (AT, BE, CH, CY, DE,
DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).
- (72) 発明者; および 添付公開書類:
(75) 発明者/出願人 (米国についてのみ): 高本良史 ー 国際調査報告書
(TAKAMOTO, Yoshifumi) [JP/JP]. 打桐竜巳 (UCHI-
GIRI, Tatsumi) [JP/JP]; 千185-8601 東京都国分寺市 2文字コード及び他の略語については、定期発行される
東恋ヶ窪一丁目280番地 株式会社 日立製作所 中央 各PCTガゼットの巻頭に掲載されている「コードと略語
のガイダンスノート」を参照。

(54) Title: STORAGE AREA NETWORK SYSTEM

(54) 発明の名称: ストレージエリアネットワークシステム



- | | |
|------------------------------|---------------------------------------|
| 101 ... LOCAL AREA NETWORK | 114 ... HOST ADAPTER |
| 102 ... HOST 1 | 115 ... HOST ADAPTER |
| 103 ... HOST 2 | 116 ... HOST ADAPTER |
| 104 ... STORAGE AREA NETWORK | 117 ... CACHE |
| 105 ... DISK CONTROL UNIT | 118 ... CACHE |
| 106 ... DISK CONTROL UNIT | 119 ... CACHE |
| 107 ... DISK CONTROL UNIT | 120 ... DISK ADAPTER |
| 111 ... DISK | 121 ... DISK ADAPTER |
| 112 ... DISK | 122 ... DISK ADAPTER |
| 113 ... DISK | A ... DISK SYSTEM I/O CONTROL PROGRAM |

(57) Abstract: The reliability of a remote copy system connected through a dedicated interface decreases until a faulty disk system recovers from trouble. To solve the problem, an auxiliary disk system connected with a storage area network is substituted for the primary disk system if it is out of order. A reliable copy system can thus be obtained.

WO 01/29647 A1

[続葉有]



(57) 要約:

専用インターフェースで接続されたリモートコピーシステムでは、障害を起こしたディスクシステムが復旧するまでの間、信頼性が低下する。

そこで、プライマリディスクシステム障害時に、ストレージエリアネットワークに接続された予備ディスクシステムと交換する。

その結果、高い信頼性を持つリモートコピーシステムが得られる。

明 細 書

ストレージエリアネットワークシステム

技術分野

本発明は記憶装置であるディスクシステムがストレージエリアネットワークに接続された場合のリモートコピーを行うストレージエリアネットワークに関する。

背景技術

計算機のデータを記憶する装置として、コストパフォーマンスが高い磁気ディスク装置が一般的に使用される。磁気ディスクは2.5インチや3.5インチ程度の複数の磁気円盤と、各磁気円盤の両面に設けられた磁気ヘッドによりデータが読み書きされる機構である。

磁気ディスクは、機械的動作を伴うため処理の時間は10m秒程度とプロセッサの処理速度などと比較すると遅い。プロセッサが高速化されても、ディスクが高速化されないとシステム全体の性能が向上しないケースも多い。この問題を解決する手段としてディスクアレイがある。ディスクアレイは、「Understanding I/O Subsystems First Edition」, W. David Schwaderer, Andrew W. Wilson, Jr. 著, page 271～291に述べられているように、複数のドライブにデータを分散して配置するとともに冗長データもドライブに格納することで、性能と信頼性を向上させる方式である。大規模なシステムでは、要求される全ドライブ容量も大きくなり、性能と信頼性が同時に要求されるためディスクアレイが使用される。

ディスクアレイが装置単体の信頼性を高めているのに対し、複数のディスクアレイを使用し広域に渡って高い信頼性を得る方法が、USP 5 8 7 0 5 3 7にて述べられている。USP 5 8 7 0 5 3 7では、2台のディスク制御装置は、メインフレーム専用光インターフェース（ESCON）で接続され、一つはプライマリディスク装置、もう一つはセカンダリディスク装置として定義される。ホストコンピュータは2台あり、一つはプライマリディスク装置とセカンダリディスク装置に接続され、もう一つはセカンダリディスク装置のみに接続されている。リモートコピーでは、プライマリディスク装置に接続されたホストコンピュータから、プライマリディスク装置に対して書き込み要求が発行されると、プライマリディスク装置はセカンダリディスク装置に、前述のESCONを介して書き込み要求を転送し、セカンダリディスク装置に同じデータが格納される。こうすることで、片方のストレージに障害が発生しても、もう片方のストレージで処理が継続される。さらに、USP 5 8 7 0 5 3 7では、リモートコピーシステムに障害が発生した場合の動作が記述されており、プライマリディスク装置に障害が発生した場合、セカンダリディスク装置にホストコンピュータからのパスを切りかえることで処理が継続されること、およびプライマリディスク装置が障害から復旧した場合、セカンダリディスク装置とプライマリディスク装置が入れ替えられることが述べられている。

発明の開示

ディスクアレイが高速化に対応し、ディスクアレイとホストコンピュータを接続するインタフェースとしてファイバチャネルが有望視されている。ファイバチャネルは、従来一般的に使用されていたSCS

I (Small Computer System Interface) の欠点である性能と接続性に優れている。特に、接続性はSCSIが数10mまでしか接続距離が伸ばせないのに対し、ファイバチャネルは数Kmまで伸ばすことができ、また、接続可能な装置数も数倍と多い。ファイバチャネルは広範囲のデバイスやホストコンピュータを接続できることで、ホストコンピュータ間のデータ通信に用いられているローカルエリアネットワークに対応し、ストレージエリアネットワークと呼ばれることもある。ファイバチャネルは、規格化されており、規格に合ったデバイスやホストコンピュータであれば、ストレージエリアネットワークに接続することができる。例えば、ファイバチャネルインタフェースを持つ複数のディスクアレイと複数のホストコンピュータを接続すること等が可能となる。

しかし、前述のUSP 5 8 7 0 5 3 7の場合、ディスク装置間は専用インタフェースで接続される形態であるため、ストレージエリアネットワークを介したリモートコピーには対応していない。また、プライマリディスク装置またはセカンダリディスク装置に障害が発生した場合、USP 5 8 7 0 5 3 7では障害が発生した装置を修復するまでリモートコピーのペアが復旧しない。これは、ディスク装置間を専用インターフェースで接続しているためである。リモートコピーのペアとなる装置を専用インターフェースで接続した場合、専用インターフェースで接続されたディスク装置間でしかデータ転送ができないためである。また、USP 5 8 7 0 5 3 7では、プライマリディスク装置に障害が発生した場合、ホストコンピュータが入出力先をセカンダリディスク装置に切りかえることでホストコンピュータの処理を継続しているが、ホストコンピュータ側で切り替え処理が必要となり、入出力のオーバーヘッドが増加する問題がある。さらに、USP 5 8 7 0 5

37では、ホストコンピュータとディスク装置の接続パスとリモートコピーのためのディスク装置間の接続パスが異なっている。そのため、リモートコピー時にパスを渡るオーバーヘッドが増加する。

なお、USP 5 8 7 0 5 3 7にはプライマリディスク装置が復旧した場合の記載がない。しかし、プライマリサイトとセカンダリサイトが離れている点、及び、セカンダリホストコンピュータがセカンダリディスク装置にしかアクセスできないことから、障害復旧時には、プライマリホストコンピュータはプライマリディスク装置に入出力先を切替えるものと思われる。

上記問題を解決するために、ストレージエリアネットワークを介してリモートコピーを行うようにするとともに、リモートコピーに関する記憶装置に障害が発生した場合、障害発生した記憶装置の代わりとしてストレージエリアネットワークに接続された予備記憶装置を割当てる。また記憶装置であるディスク装置に、ホストコンピュータとの接続を制御するホストコンピュータアダプタを2つ設け、プライマリディスク装置またはセカンダリディスク装置に障害が発生した場合、2つのホストコンピュータアダプタの内ひとつを障害が発生したディスク装置の装置IDに変更することで、ホストコンピュータを変更することなく処理を継続させる。さらに、ホストコンピュータから転送されたコマンドを、変更を加えることなくストレージエリアネットワークへ転送し直すプライマリディスク装置と、自装置IDとは異なるコマンドを受け付けセカンダリディスク装置を設ける。

図面の簡単な説明

図1は、本発明の実施例1における全体構成図を示す図である。

図2は、ディスクシステムの構成図を示す図である。

図 3 は、ホストコンピュータプログラムの階層を示す図である。

図 4 は、正常時のリモートコピー構成を示す図である。

図 5 は、障害時のリモートコピー構成を示す図である。

図 6 は、ホストコンピュータリモートコピー管理テーブルを示す図である。

図 7 は、ディスクシステム制御プログラムのフロー示す図である。

図 8 は、ストレージエリアネットワークと装置の距離の関係を示す図である。

図 9 は、セカンダリディスクシステム選択処理フローを示す図である。

図 10 は、正常時のリモートコピー構成を示す図である。

図 11 は、障害時のリモートコピー構成を示す図である。

図 12 は、ホストコンピュータアダプタリモートコピー管理テーブルを示す図である。

図 13 は、ホストコンピュータアダプタプライマリ移行処理フローを示す図である。

図 14 は、プライマリリモートコピー処理フローを示す図である。

図 15 は、セカンダリリモートコピー処理フローを示す図である。

図 16 は、ホストコンピュータアダプタの構成を示す図である。

図 17 は、リモートコピーシステム構成例を示す図である。

図 18 は、コマンドパケット構成を示す図である。

図 19 は、リモートコピーの処理フローを示す図である。

図 20 は、性能比較を示す図である。

発明を実施するための最良の形態

[実施例 1]

図 1 は、本発明によるストレージエリアネットワーク環境におけるリモートコピーシステムの概略構成図を示している。101 はローカルエリアネットワークであり、102, 103 はそれぞれホストコンピュータである。105, 106, 107 はそれぞれディスクシステムである。ホストコンピュータ 102, 103 とディスクシステム 105, 106, 107 はストレージエリアネットワーク 104 に接続される。主に、ローカルエリアネットワーク 101 はホストコンピュータ間の通信に利用され、ストレージエリアネットワーク 104 はディスクシステム 105, 106, 107 とホストコンピュータ 102, 103 間、またはディスクシステム 105, 106, 107 間のデータ通信に用いられる。一般に、ローカルネットワーク 101 よりストレージエリアネットワーク 104 の方がデータ転送性能が高い。従って、ストレージエリアネットワークの方が大規模なデータ転送に向く。ホストコンピュータ 102, 103 内のディスクシステム入出力制御プログラムは、ホストコンピュータ 102, 103 内で実行されているアプリケーションプログラム（図示せず）の入出力要求をディスクシステムに発行する機能を有する。ディスクシステム 105, 106, 107 内は、ディスク制御装置 108, 109, 110 とディスク 111, 112, 113 から構成される。ディスク制御装置 108, 109, 110 は、ホストコンピュータ 102, 103 から発行された入出力要求を解釈／実行するハードウェアであり、ディスク 111, 112, 113 はホストコンピュータ 102, 103 から転送されたデータを格納する機能を有する。ディスク制御装置 108, 109, 110 内はホストコンピュータアダプタ 114, 115, 116 とキャッシュ 117, 118, 119 とディスクアダプタ 120, 121, 122 から構成される。ホストコンピュータアダプタ 114

， 1 1 5， 1 1 6 はホストコンピュータ 1 0 2， 1 0 3 から発行されたコマンドを受信し解釈する機能を有し、ディスクアダプタ 1 2 0， 1 2 1， 1 2 2 はホストコンピュータアダプタ 1 1 4， 1 1 5， 1 1 6 の解釈結果に基づきディスク 1 1 1， 1 1 2， 1 1 3 に対して入出力を実行する機能を有する。キャッシュ 1 1 7， 1 1 8， 1 1 9 は、ディスク 1 1 1， 1 1 2， 1 1 3 から読み込まれたデータを一時的に格納したり、ホストコンピュータ 1 0 2， 1 0 3 から転送された書き込みデータを一時的に格納する領域である。再度同一データの読み込みがホストコンピュータ 1 0 2， 1 0 3 から要求された場合に、キャッシュ 1 1 7， 1 1 8， 1 1 9 からホストコンピュータ 1 0 2， 1 0 3 へデータを返送することで入出力レスポンスを高速化することが可能となる。またホストコンピュータ 1 0 2， 1 0 3 へはキャッシュ 1 1 7， 1 1 8， 1 1 9 へデータが格納された時点で書き込み完了報告を行うため、書き込みを高速に行ったように見せることができる。図 1 では、ディスクシステム 1 0 5 がリモートコピーにおけるプライマリディスクシステムであり、ディスクシステム 1 0 6 がセカンダリディスクシステムである。また、ディスクシステム 1 0 7 は、予備ディスクシステムであり、プライマリディスクシステム 1 0 5 またはセカンダリディスクシステム 1 0 6 のいずれかに障害が発生した場合の交替用である。ホストコンピュータ 1 0 2， 1 0 3 はリモートコピーシステムが正常な場合は、プライマリディスクシステム 1 0 5 に対して、入出力要求を発行する。ホストコンピュータ 1 0 2， 1 0 3 からプライマリディスクシステム 1 0 5 に対して書き込み要求が発行されると、プライマリディスクシステムはストレージエリアネットワーク 1 0 4 を介して書き込みデータをセカンダリディスクシステム 1 0 6 に転送する。その結果、プライマリディスクシステム 1 0 5 とセカンダリ

ディスクシステム 106 には、同じデータが格納される。プライマリディスクシステム 105 からセカンダリディスクシステム 106 へのデータ転送はホストコンピュータ 102, 103 が意識すること無く実行される。

図 2 は、ディスクシステムの詳細な構成を示している。ディスクシステム 105, 106, 107 は、ディスク制御装置 201 とディスク 207, 208, 209, 210 から構成される。ディスク制御装置内はホストコンピュータアダプタ 202, 203 とディスクキャッシュ 204 とディスクアダプタ 205, 206 から構成される。ホストコンピュータアダプタ 202, 203 とディスクキャッシュ 204 とディスクアダプタ 205, 206 はバス 211 で接続され、それぞれの構成要素間でバス 211 を介して通信することができる。ホストコンピュータアダプタ 202, 203 はホストコンピュータ 102, 103 から発行されたコマンドを受信し解釈する機能を有し、一つのディスク制御装置 201 内に複数設けることができる。ディスクアダプタ 205, 206 はホストコンピュータアダプタ 202, 203 の解釈結果に基づきディスク 207, 208, 209, 210 に対して入出力を実行する機能を有するが、ホストコンピュータアダプタ 202, 203 と同様に一つのディスク制御装置 201 内に複数設けることができる。ホストコンピュータアダプタ 202, 203 やディスクアダプタ 205, 206 を複数設けることによって、処理の負荷分散が可能になり、また信頼性も向上する

図 3 は、ホストコンピュータ 102, 103 内のプログラムの構成を示している。最も上位に位置するプログラムはアプリケーションプログラム 301 である。アプリケーションプログラム 301 は、通常はユーザが記述したプログラムであり、ディスクシステム 105, 1

06, 107に対する入出力要求の発端となる処理を行う。ミドルプログラムは、データベースなどのアプリケーションプログラム301とオペレーティングシステム303の間に位置する。オペレーティングシステム303は、ホストコンピュータの管理やハードウェアの制御を行うプログラムであり、ホストコンピュータ上のプログラム階層では最も下位に位置する。オペレーティングシステム303内には、ファイルシステム304とディスク入出力制御プログラム305がある。

図4, 5は、本発明の特徴の一つを示している。図4は、ストレージエリアネットワーク401環境の正常時のリモートコピーの形態を示している。ディスクシステム1(402)はプライマリディスクシステム、ディスクシステム2(403)はセカンダリディスクシステム、ディスクシステム3(404)は予備ディスクとして定義されている。図5は、プライマリディスクシステム502に障害が発生した場合に、本発明によってリモートコピーの定義がどのように変更されるかを示している。本発明では、プライマリディスクシステム502に障害が発生すると、セカンダリシステムとして定義されていたディスクシステムをプライマリディスクシステムとして定義し直し、また、予備ディスクシステム504をセカンダリディスクシステムとして定義する。こうすることで、障害が発生したディスクシステムが修復されるまで、2重化ができない従来システムと異なり、高い信頼性を得ることができるようになる。

本実施例では、図4から図5の形態への変更は、ホストコンピュータが指示する。ホストコンピュータ内には、図6に示すリモートコピーのための、ホストコンピュータリモートコピー管理テーブルが格納されている。カラム601は、リモートコピーに関するディスクシ

システムのデバイス識別子を示している。カラム 6 0 2 は、ホストコンピュータとの距離を示している。カラム 6 0 3 はプライマリディスクシステムとの距離を示している。カラム 6 0 4 は、現在の属性を示している。例えば、デバイス ID が 0 1 のディスクシステムはプライマリディスクシステムとして定義されており、デバイス ID が 0 3 のディスクシステムは、デバイス ID が 0 1 のディスクシステムのセカンダリディスクシステムと定義されている。その他のディスクシステムは予備ディスクシステムとして定義されている。カラム 6 0 2, 6 0 3 は、主に予備ディスクをプライマリディスクシステムやセカンダリディスクシステムに割当てるときに参照される。リモートコピーは、遠距離に配置されたディスクシステム間で同じデータを保持することに大きな意味がある。ディスクシステムの構成部品やファームウェアの誤動作によるディスクシステム障害を防ぐことが目的であれば、ディスクシステムの二重化で対応可能である。しかし、大規模な電源障害や災害などによりシステムが動作不可能になることを防ぐには、二重化されたディスクシステムをできるだけ遠距離においた方が信頼性を高くできる。これは、同時に障害が発生する確率を小さくできるためである。従来のリモートコピーシステムでは、リモートコピーのために専用のインターフェースケーブルを設置していたため、頻繁にシステムを変更することができなかった。そのため、カラム 6 0 2, 6 0 3 の距離に関する情報は必要ない。しかし、本発明によるストレージエリアネットワーク環境のリモートコピーシステムでは、動的に構成を変更することができる。そのため、リモートコピーシステムに新しく構成を追加する場合は、ディスクシステムがどこに配置されているかを把握しておき、適切なディスクシステムを選択する必要がある。

図 7 は、ホストコンピュータ 102, 103 内のディスクシステム入出力制御プログラムのフローを示している。ステップ 701 は、ディスクシステムへ入出力要求を発行する。ステップ 702 は、ステップ 701 で発行された入出力要求の終了判定を行う。正常に終了した場合は、本プログラムを終了する。正常にできなかった場合は、ステップ 703 に進む。ステップ 703 は、エラーが発生した入出力システムがリモートコピーとして定義されているかどうか判断する。これは、図 6 のホストコンピュータリモートコピー管理テーブルを参照することで判別できる。リモートコピーとして定義されていない入出力システムの場合は、ステップ 704 に進み、ファイルシステム（図 3 の 304）に対してエラーを報告する（ステップ 704）。一方、エラーが発生したシステムがリモートコピーシステムの構成要素の一つであった場合は、ステップ 705 に進む。ステップ 705 は、エラーが発生した入出力システムがプライマリディスクシステム化どうかを判別する。これは、図 6 のホストコンピュータリモートコピー管理テーブルを参照することで判別できる。プライマリディスクシステムがエラーを起こした場合はステップ 706 に進み、そうでない場合は 707 に進む。ステップ 706 は、予備ディスクシステムをセカンダリディスクシステムとして定義する。本ステップでは、図 6 のホストコンピュータリモートコピー管理テーブルの更新と、変更を各ディスクシステムへ伝える。ステップ 708 では、これまでセカンダリディスクシステムだったディスクシステムをプライマリディスクシステムに定義し直す。ステップ 709 では、エラーを起こした入出力要求を再発行する。一方、ステップ 707 は、セカンダリディスクシステムが障害を起こした場合に実行され、ステップ 706 と同じように、セカンダリディスクシステムを割当ててることを意味する。こうすることで、

プライマリディスクシステムにエラーが発生しても、即座にリモートコピーシステム構成を再構築することができる。

図 8, 9 は、本発明におけるリモートコピーシステムへの、予備ディスクシステム追加方法を示している。図 8 の 801 はホストコンピュータ、803 はプライマリディスクシステム、802, 804 は予備ディスクシステムである。例えば、新たにセカンダリシステムを追加する場合、802 および 804 のいずれかを選択することができる。リモートコピーシステムを効果的に運用するためには、リモートコピーシステムの各構成要素が、どこに配置されているかが重要な指標になる。リモートコピーは、遠距離に配置されたディスクシステム間で同じデータを保持することに大きな意味がある。ディスクシステムの構成部品やファームウェアの誤動作によるディスクシステム障害を防ぐことが目的であれば、距離を意識しないディスクシステムの二重化で対応可能である。しかし、電源障害や災害などによりシステムが動作不可能になることを防ぐには、二重化されたディスクシステムをできるだけ遠距離においた方が信頼性を高くできる。これは、遠距離に装置を配置した方が、電源障害や災害時に、同時障害が発生する確率を小さくできるためである。図 8 では、ホストコンピュータ 801 と予備ディスクシステム 802 の距離 (805) よりも、ホストコンピュータ 801 と予備ディスクシステム 804 の距離 (807) の方が遠距離である。一方、予備ディスクシステム 802 とプライマリディスクシステム 803 の距離 (806) と、予備ディスクシステム 804 とプライマリディスクシステム 803 の距離 (808) はほぼ同じである。遠隔地にディスクシステムを置くことで障害に備えることに重点を置けば、図 8 の例では予備ディスクシステム 804 をセカンダリディスクシステムに割当てることがより高い信頼性を得ることが

できる。図7は、本実施例における、予備ディスクシステムの選択方法を示したフローである。ステップ901では、図6のホストコンピュータリモートコピー管理テーブルから予備ディスクとして定義されたディスクシステムを抽出する。ステップ902では、ステップ901で抽出された予備ディスクシステムのホストコンピュータとの距離とプライマリディスクシステムとの距離の積を求め、この数値が最も大きい予備ディスクシステムを選択する。このフローは、図9のステップ706および707に適用することができる。このようにして、予備ディスクを選択することで、効果的なリモートコピーシステムを構築することができるようになる。本実施例では、ディスクシステム間の距離だけでなく、ホストシステムとディスクシステム間の距離を離すことができる。従来のリモートコピーシステムのようにディスクシステム間だけが遠距離に配置されるだけでなく、ディスクシステムとホスト間も遠距離に配置できるため、従来のリモートコピーシステムに比べ、より信頼性が高いリモートコピーシステムを提供できる。本発明のポイントの一つは、ストレージエリアネットワーク環境においてリモートコピーシステムに新しく構成を追加する場合は、距離を考慮したほうがより効果的なシステムを構築可能なことである。こうすることで、電源障害や災害などによりシステムが動作不可能になる確率を小さくできるため、より高い信頼性を得ることができるようになる。

図10、11は、本発明の特徴の一つを示している。本発明の特徴の一つは、プライマリシステムに障害が発生しても、ホストコンピュータ上のいかなるプログラムも変更すること無く運用が継続できることである。図10は、本発明における正常時のリモートコピー構成を示している。1001はプライマリディスクシステムであり、100

2はセカンダリディスクシステムである。プライマリディスクシステム1001内にはホストコンピュータアダプタ1003があり、またセカンダリディスクアレイ1002内には2つのホストコンピュータアダプタ1004, 1005がある。それぞれのホストコンピュータアダプタ1003, 1004, 1005はストレージエリアネットワーク1010に接続されている。ホストコンピュータアダプタ1003, 1005は、それぞれストレージエリアネットワーク1010内で固有の装置IDが付けられており、ホストコンピュータアダプタ1003は装置ID1、ホストコンピュータアダプタ1005は装置ID2である。ホストコンピュータアダプタ1004は、正常時には無効となっている。ホストコンピュータ（図示せず）から、プライマリディスクシステム1001に対して書込み要求が発生すると、ホストコンピュータアダプタ1003は、ストレージエリアネットワーク1010を介してデータをホストコンピュータアダプタ1005に転送することでリモートコピーを行う。一方、図11はプライマリディスクアレイに障害が発生した図を示している。この場合、前述の通り、従来セカンダリディスクシステムとして動作していたディスクシステムはプライマリディスクシステムに属性が変更される。障害が発生したディスクシステム1101のホストコンピュータアダプタ1103が持っていた装置IDを、プライマリディスクシステム1102内のホストコンピュータアダプタ1104が引き継ぐ。その一方で、障害が発生したディスクシステム1101内のホストコンピュータアダプタ1103は無効化する。こうすることで、ホストコンピュータ（図示せず）から発行された入出力要求は、新しいプライマリディスクシステム1102で処理可能となり、ホストコンピュータ内のさまざまなプログラムを変更することなく処理を継続できる。

上記処理を実現するために、各ディスクシステム内には図 1 2 に示す、ホストコンピュータアダプタリモートコピー管理テーブルを格納してある。カラム 1 2 0 1 はデバイス（装置）ID であり、カラム 1 2 0 2 はリモートコピーシステムの属性であり、カラム 1 2 0 3 は、リモートコピー対象ボリュームである。単一のディスクシステムでも、複数のボリュームを定義できる装置が一般的であり、ユーザはボリュームの用途に応じてリモートコピーをするか否かを選択することができる。このために、カラム 1 2 0 3 の対象ボリュームが必要となる。

図 1 3 は、セカンダリディスクシステム内のホストコンピュータアダプタで実行されるプライマリ移行処理フローを示している。ステップ 1 3 0 1 は、プライマリディスクシステムのデバイス ID を取得する。これは図 1 2 の、ホストコンピュータアダプタリモートコピー管理テーブルを参照することで取得することができる。ステップ 1 3 0 2 は、セカンダリディスクシステム内の無効化されているホストコンピュータアダプタのデバイス ID を、ステップ 1 3 0 1 で取得したデバイス ID に変更する。ステップ 1 3 0 3 はセカンダリディスクシステム内の無効化されているホストコンピュータアダプタをプライマリディスクシステムとする。ステップ 1 3 0 4 は、セカンダリディスクシステム内の無効化されているホストコンピュータアダプタを有効にする。こうすることで、プライマリディスクシステムに障害が発生しても、ホストコンピュータのハードやソフトを変更する必要はなく、かつホストコンピュータ上のプログラムを停止させる必要もなくなる。

図 1 4 から図 2 0 は、本発明の特徴の一つであるリモートコピーの高速化手法を示している。ストレージエリアネットワーク環境におけるリモートコピーは、従来のリモートコピーと構成が大きく異なって

いる。例えば、従来のリモートコピーは、プライマリディスクシステムとセカンダリディスクシステム間のデータ転送は専用のケーブルを用いて行われていた。そのため、ホストコンピュータとプライマリディスクシステム間、およびプライマリディスクシステムとセカンダリディスクシステム間は別のケーブルで接続される形態である。しかし、ストレージエリアネットワークはホストコンピュータ、プライマリディスクシステム、セカンダリディスクシステムが同一のケーブルで接続される。従来のリモートコピーでは、異なる2本のケーブルでリモートコピーシステムを構築していたため、プライマリディスクシステム内には、2つのホストコンピュータアダプタが必要である。この2つのホストコンピュータアダプタ間での通信オーバーヘッドによりリモートコピーの性能が劣化していた。図14から図20では、単一のホストコンピュータアダプタでリモートコピーを実現し、かつさらなる高速化について述べる。

図14はプライマリディスクシステム側のホストコンピュータアダプタ内の処理フローを示している。ステップ1401は、ホストコンピュータからストレージエリアネットワークを介してパケット形式で転送された入出力要求を受信する。ステップ1402では要求コマンドが書込み要求かどうか判断する。書込み要求であればステップ1403に進み、そうでなければステップ1405に進む。ステップ1405は、読込み要求を実行し処理を終了する。ステップ1403では、ステップ1401で受信したパケットをストレージエリアネットワークに転送し直す。これは、後で述べるセカンダリディスクシステムが受信することになる。ここでの特徴は、プライマリディスクシステムが受信したパケットをそのままの形式で再送することにある。こうすることで、コマンドの解析やパケット形式の変更といったオーバヘ

ッドを無くすことができる。ステップ 1 4 0 4 は、プライマリディスクシステム内でホストコンピュータの書込みコマンドを実行する。ステップ 1 4 0 6 は、書込みコマンドがリモートコピーの対象ボリュームに対して行われたかどうかを判断する。これは、図 1 2 のホストコンピュータアダプタリモートコピー管理テーブルを参照することで判断できる。リモートコピーの対象ボリュームに対する書込みであればステップ 1 4 0 7 に進み、そうでなければ終了する。ステップ 1 4 0 7 でセカンダリディスクシステムからのリモートコピー終了を待ち、その後処理は終了する。

図 1 5 は、セカンダリディスクシステム側のリモートコピー処理フローを示している。ステップ 1 5 0 1 は、ストレージエリアネットワークを介してパケット形式で転送された入出力要求を受信する。ステップ 1 5 0 2 では、転送された入出力要求がプライマリディスクアレイから転送された書込みコマンドかどうかを判断する。これは、図 1 2 のホストコンピュータアダプタリモートコピー管理テーブルを参照することで判断できる。ここでの特徴は、受信したパケットは、前述の通りプライマリディスクシステムが受信したパケットをそのまま転送し直しているため、パケット内はプライマリディスクシステムに対する情報がはいっている。そのため、ステップ 1 5 0 2 では、プライマリディスクシステムに対する要求を、セカンダリディスクシステムが受信することを意味している。ステップ 1 5 0 4 は受信したコマンドがリモートコピー対象ボリュームかどうか判断する。これは、図 1 2 のホストコンピュータアダプタリモートコピー管理テーブルを参照することで判断できる。リモートコピーの対象ボリュームに対する書込みであればステップ 1 5 0 5 に進み、そうでなければ終了する。ステップ 1 5 0 5 では、受信したコマンドに従がって書き込みを実行する。

ステップ 1507では、セカンダリディスクシステム内の書込み処理が終了したことを、プライマリディスクシステムに通知し、終了する。ステップ 1503, 1506は、プライマリディスクシステム以外の装置から転送された通常の入出力を実行し終了する。

図 16, 17, 18は、図 15におけるステップ 1501およびステップ 1502を、さらに詳細に述べた図である。図 16は、ホストコンピュータアダプタ 1601をさらに詳細化した図である。1602はインターフェース制御 LSI であり、1603は制御プロセッサであり、1604はメモリであり、1605はバス制御 LSI である。インターフェース LSI (1602)は、ストレージエリアネットワークとの通信を制御する LSI である。主に、ストレージエリアネットワークの通信プロトコル制御を行う。制御プロセッサ 1603は、ホストコンピュータアダプタ 1601の主要な制御を行い、実行すべきプログラムはメモリ 1604に格納されている。バス制御 LSI は、ストレージシステム内のバス (図 2の 211)を制御する LSI である。図 15のステップ 1501と 1502は、主にインターフェース制御 LSI 内の処理を示している。図 15のステップ 1501, 1502以外のステップは、全て制御プロセッサ 1603によって実行される。

図 17は、図 14および図 15で述べたリモートコピーの構成例を示している。ホストコンピュータ 1701とプライマリディスクシステム 1703とセカンダリディスクシステム 1704はストレージエリアネットワーク 1702に接続されている。ホストコンピュータ 1701の装置 ID は 0 であり、プライマリディスクシステム 1703内のホストコンピュータアダプタの装置 ID は 1 であり、セカンダリディスクシステム 1704内のホストコンピュータアダプタの装置 I

Dは2とする。ストレージエリアネットワーク1702内は図18に示すようなパケット形式でデータが転送される。1801はフレームの先頭を示すフィールドであり、1802はフレームヘッダであり、1803はI/Oコマンドやデータが格納され、1804はパケット内のデータに誤りが無いかどうかをチェックするためのコードが格納されており、1805はフレームの終わりを示すフィールドである。フレームヘッダ内はさらに、送信先デバイスID(1806)、送信元デバイスID(1807)、および制御フィールドから構成される。例えば、図17において、ホストコンピュータ1701からプライマリディスクシステム1703にパケットが転送される際には、送信先デバイスIDには1が格納され、送信元デバイスIDには0が格納される。入出力コマンドは1803内に格納されている。

図19は、発明におけるリモートコピーの一実施例の動作例を示している。1901はホストコンピュータの動作を示しており、1902はプライマリディスクシステムの動作を示しており、1903はセカンダリディスクシステムの動作を示している。最初にホストコンピュータから書き込みコマンドが発行される(1904)。このコマンドはプライマリディスクシステムが受信する(1905)。送信先IDはプライマリディスクシステムの1が格納され、送信元IDはホストコンピュータを示す0が格納されている。このコマンドは直ちにセカンダリディスクシステムに転送され、セカンダリディスクシステムが受信する(1907)。プライマリディスクシステムはパケットの送信が終了すると書き込みコマンドの実行を開始する(1908)。セカンダリディスクシステムが受信したパケットは、1905で受信した内容と同じである。そのため、送信先IDはプライマリディスクシステムを示す1が格納され、送信元IDはホストコンピュータを示

す0が格納されている。この packets をセカンダリシステムが受けて実行すべきコマンドかどうか判断し（1909）、実行すべきコマンドであればそのコマンドを実行する（1910）。セカンダリディスクシステムの手込み処理が終了すると、終了報告をプライマリディスクシステムに転送する（1911）。プライマリディスクシステムは、セカンダリディスクシステムからの終了通知を待つ（1912）、プライマリとセカンダリの両方で書き込み処理が完了したことを確かめてホストコンピュータに書き込み完了通知を発行する（1913）。これにより、本発明によりリモートコピーが完了する（1914）。

図20は、本発明におけるリモートコピーの性能と従来のリモートコピーと同じフローにしたがって処理した場合の性能を比較した図である。従来は、プライマリ側で、パケット受信（2001）、コマンド解析（2002）の後、セカンダリに対してパケットを送信する（2003）。プライマリ側では、書き込み処理を行い（2004）、セカンダリ側の書き込み処理が終了するのを待つ（2005）。セカンダリ側では、パケットを受信（2007）、コマンド解析（2008）、書き込み処理（2009）の後、プライマリに終了を報告する。プライマリ側では、終了通知パケットを受信し（2005）、ホストコンピュータに対して書き込み処理の終了を報告する（2006）。これに対し、本発明における実施例では、プライマリ側ではホストコンピュータ発行のパケットを受信（2010）した後、すぐにセカンダリに対してパケットを送信する（2011）。セカンダリ側では、このパケットを受信し（2016）、コマンド解析（2017）、書き込み処理を実行（2018）する。一方、プライマリ側では、セカンダリに対してパケット送信（2011）後、コマンド解析（2012）、

書込み処理実行（２０１３）の後、セカンダリ側の書込み処理を待つて（２０１４）、ホストコンピュータに終了通知を発行する（２０１５）。本発明では、従来方式に従った処理に比べ、セカンダリへのパケット送信を早めることができるため、リモートコピーのレスポンスを高速化できる。

また、本実施例の他の効果として、従来のリモートコピーではできなかった、複数のホストコンピュータによるリモートコピーシステムの共有がある。図１に示すように、ストレージエリアネットワーク１０１には複数のホストコンピュータ（１０２，１０３）を接続することができる。複数のホストコンピュータは、互いにリモートコピーシステムを共有できる。そのため、本実施例で述べたリモートコピーシステムでは、例えばホストコンピュータ１０２に障害が発生した場合でも、ホストコンピュータ１０３が図６に示すホストコンピュータリモートコピー管理テーブルを保持しておくことで、ホストコンピュータ１０２と同じリモートコピーシステム構成のままホストコンピュータ１０３は処理を引き継ぐことができる。従来リモートコピーシステムでは、セカンダリディスクシステムに接続されたホストコンピュータはセカンダリディスクシステムしかアクセスすることができなかったが、本実施例では複数のホストコンピュータから、同じ構成のリモートコピーシステムを共有できるため、より信頼性が高いリモートコピーシステムを構築することができる。

請 求 の 範 囲

1. ストレージエリアネットワークと、該ストレージエリアネットワークに接続されたホストコンピュータと、上記ストレージエリアネットワークに接続されたプライマリ記憶装置及びセカンダリ記憶装置とを有し、上記プライマリ記憶装置は該ホストコンピュータからの書きこみ要求に応じて該書きこみ要求を実行するとともに上記セカンダリ記憶装置に書きこみコマンドを転送してリモートコピーを実行し、該プライマリ記憶装置または上記セカンダリ記憶装置のいずれかに障害が発生した場合、障害が発生した記憶装置を予備記憶装置に交換することを特徴とするストレージエリアネットワークシステム。

2. 該プライマリ記憶装置障害時に、該セカンダリ記憶装置をプライマリ記憶装置とし、該予備記憶装置をセカンダリ記憶装置とすることを特徴とする請求項1記載のストレージエリアネットワークシステム。

3. 該予備記憶装置への切替は、該ホストコンピュータが行うことを特徴とする請求項1記載のストレージエリアネットワークシステム。

4. ストレージエリアネットワークと、該ストレージエリアネットワークに接続されたプライマリ記憶装置と、該ストレージエリアネットワークに接続されたセカンダリ記憶装置及び複数の予備記憶装置とを有し、該セカンダリ記憶装置に障害が発生した場合に該複数の予備記憶装置のそれぞれについてホストコンピュータとの距離とプライマリディスクシステムとの距離の積を求め、この数値が最も大きい予備記憶装置を新たにセカンダリ記憶装置とすることを特徴とするストレージエリアネットワークシステム。

5. ストレージエリアネットワークと、該ストレージエリアネットワークに接続されたホストコンピュータと、該ストレージエリアネットワークに接続されたプライマリ記憶装置及びセカンダリ記憶装置とを有し、該プライマリ記憶装置は該ホストコンピュータの書きこみ要求に応じて該書きこみ要求を実行するとともにセカンダリ記憶装置に書きこみコマンドを転送し、該セカンダリ記憶装置は複数のホストコンピュータインタフェースを持つとともに、該プライマリ記憶装置障害時に、該プライマリ記憶装置の装置識別子を該複数のホストコンピュータインタフェースの内ひとつに割当ててことを特徴とするストレージエリアネットワークシステム。

6. 該プライマリ記憶装置障害時に、該プライマリ記憶装置のホストコンピュータインタフェースを無効にすることを特徴とする請求項5記載のストレージエリアネットワークシステム。

7. ホストコンピュータとセカンダリ記憶装置が接続されたストレージエリアネットワークに接続され、該ホストコンピュータの書きこみ要求に応じて該書きこみ要求を実行するとともに上記セカンダリ記憶装置に書きこみ要求のコマンドを転送するリモートコピーをする際に、該ホストコンピュータからのコマンドをそのままストレージエリアネットワークを介して上記セカンダリ記憶装置に転送し直すことを特徴とするプライマリ記憶装置。

8. ホストコンピュータとプライマリ記憶装置が接続されたストレージエリアネットワークに接続され、該ホストコンピュータから上記プライマリ記憶装置に書きこみ要求のコマンドが送信されると、該コマンドを受け付けることを特徴とするセカンダリ記憶装置。

9. 第1と第2の記憶装置に接続されたストレージエリアネットワークに接続され、上記第1の記憶装置をプライマリ記憶装置と定義し

、上記第 2 の記憶装置を上記プライマリ記憶装置のリモートコピー先となるセカンダリ記憶装置と定義し、上記第 1 の記憶装置の障害を検知した場合に上記第 2 の記憶装置をプライマリ記憶装置と定義を変えることを特徴とするホストコンピュータ。

10. 請求項 9 において、上記第 1 の記憶装置の障害が復旧した場合に該第 1 の記憶装置をセカンダリ記憶装置と定義することを特徴とするホストコンピュータ。

図1

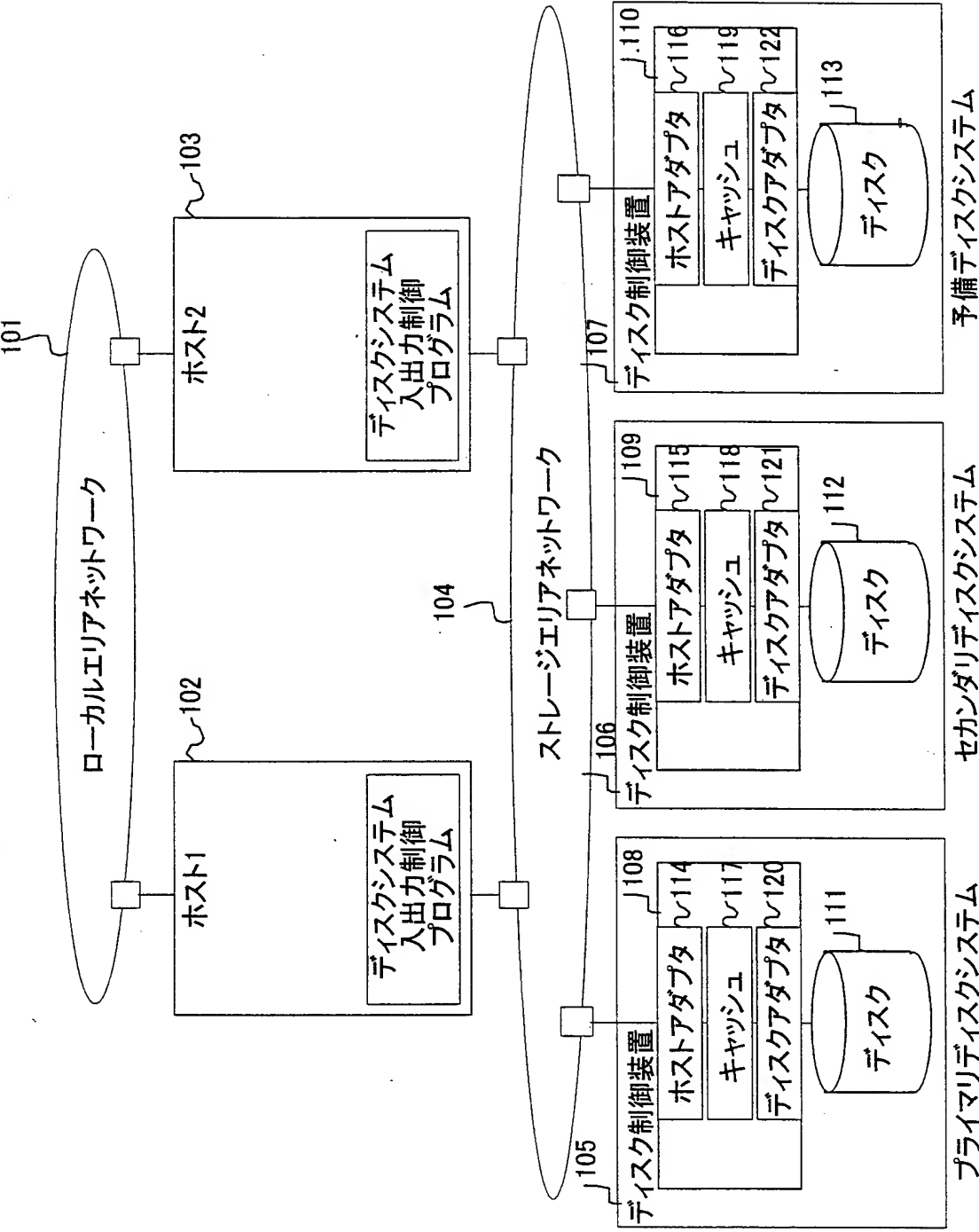


図2

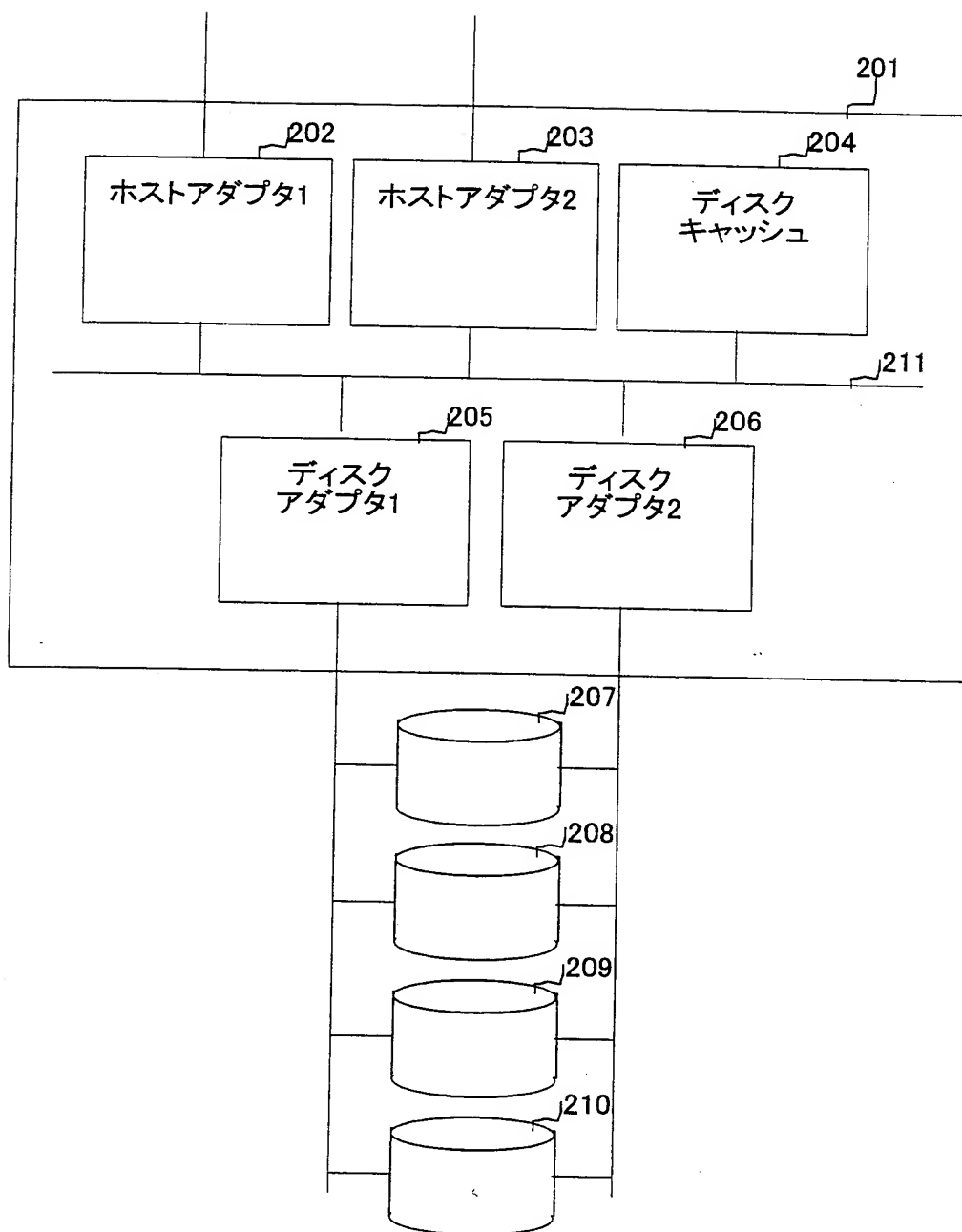


図3

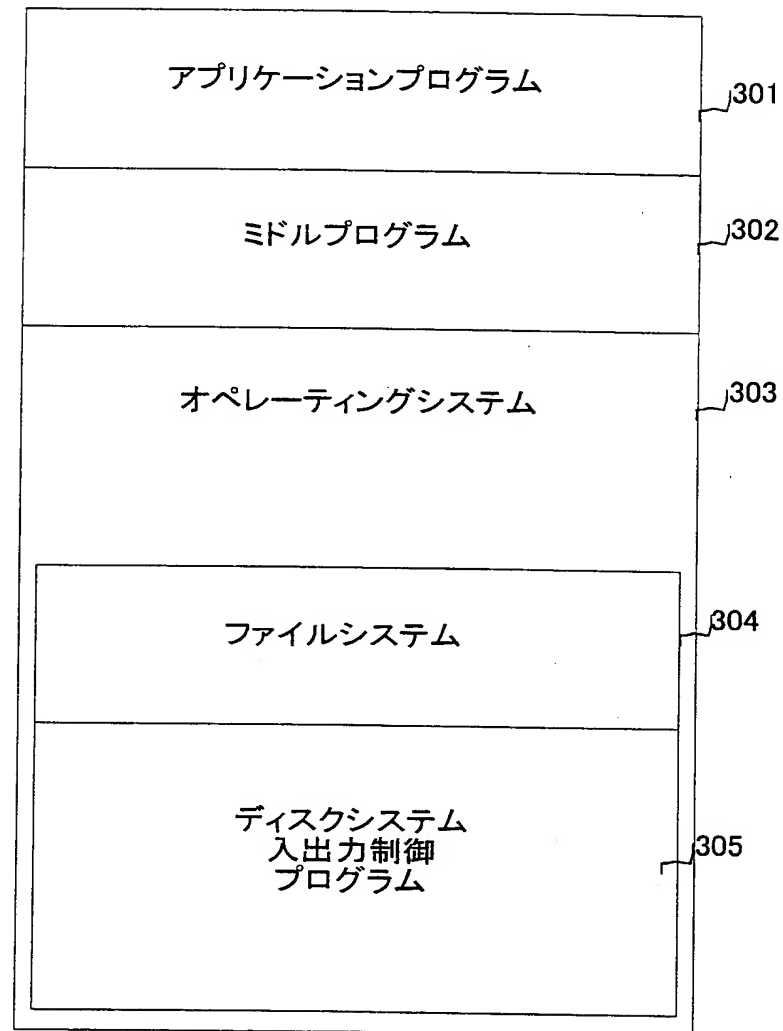
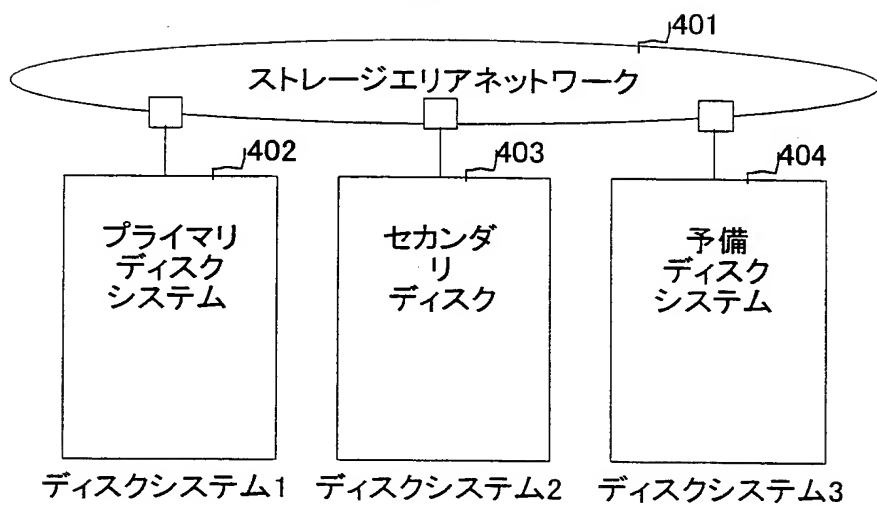


図4



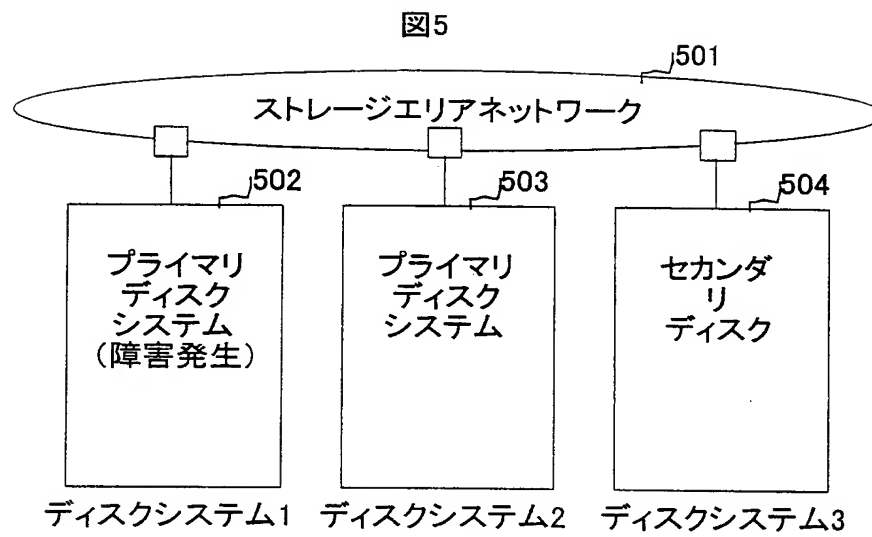


図6

ホスト
リモートコピー管理テーブル

デバイスID	ホストとの距離	プライマリ ディスクシステム との距離	属性
01	6	6	プライマリ
02	5	7	予備
03	7	5	デバイスID 01の セカンダリ
04	8	3	予備

図7

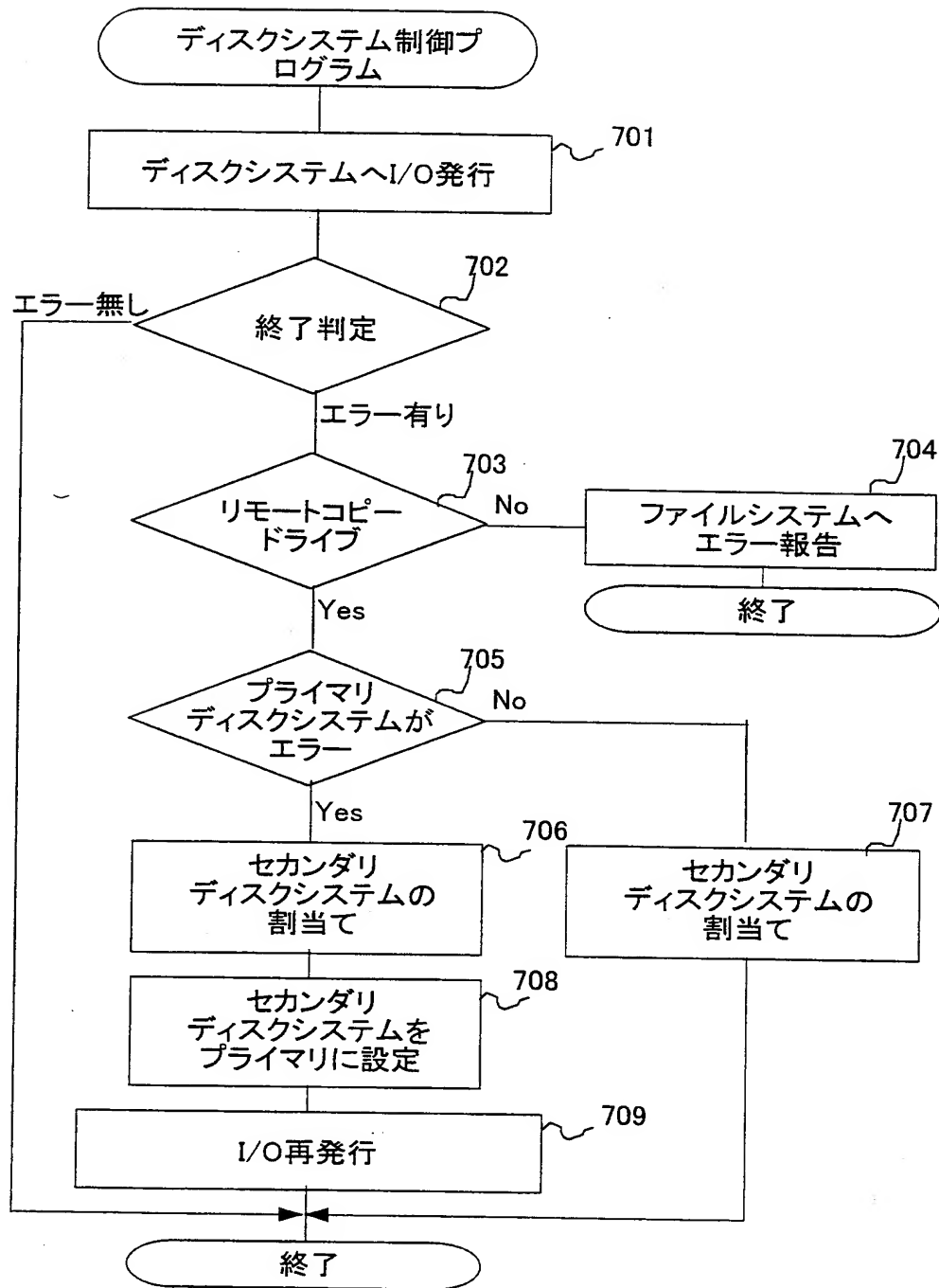


図8

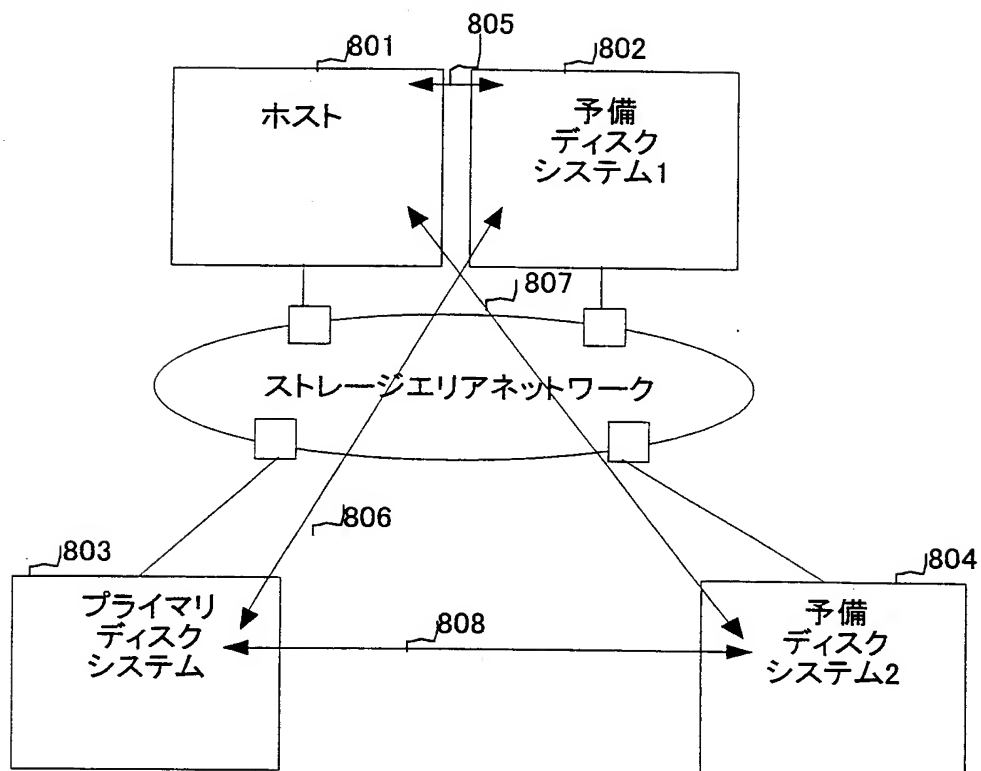


図9

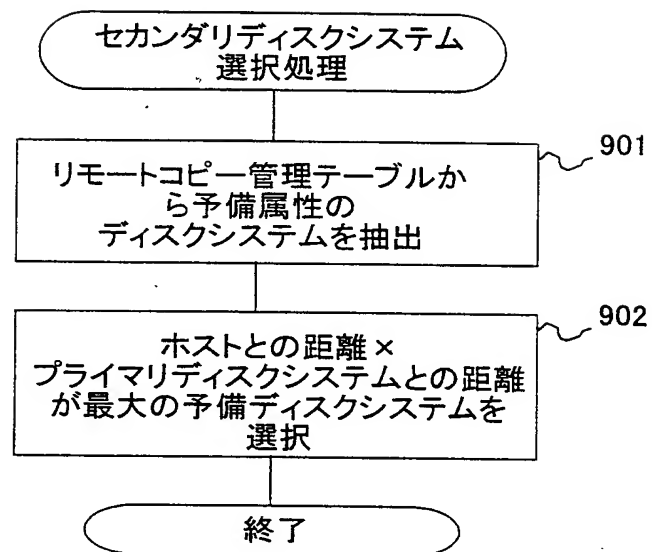
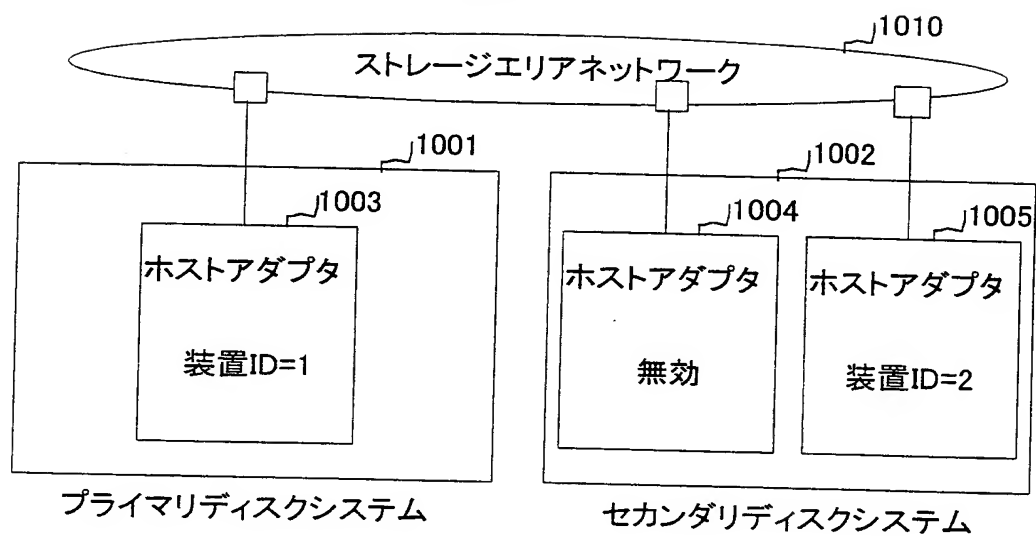


図10



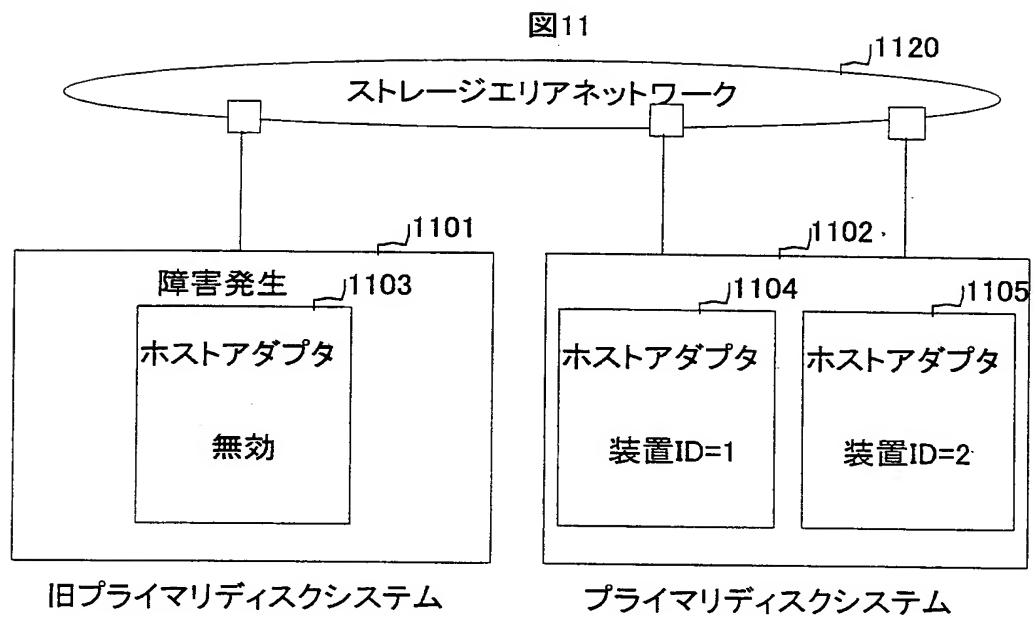


図12

ホストアダプタ
リモートコピー管理テーブル

デバイスID 1201	属性 1202	対象ボリューム 1203
01	プライマリ	VOL1,VOL2,VOL3
02	予備	
03	デバイスID 01の セカンダリ	
04	予備	

図13

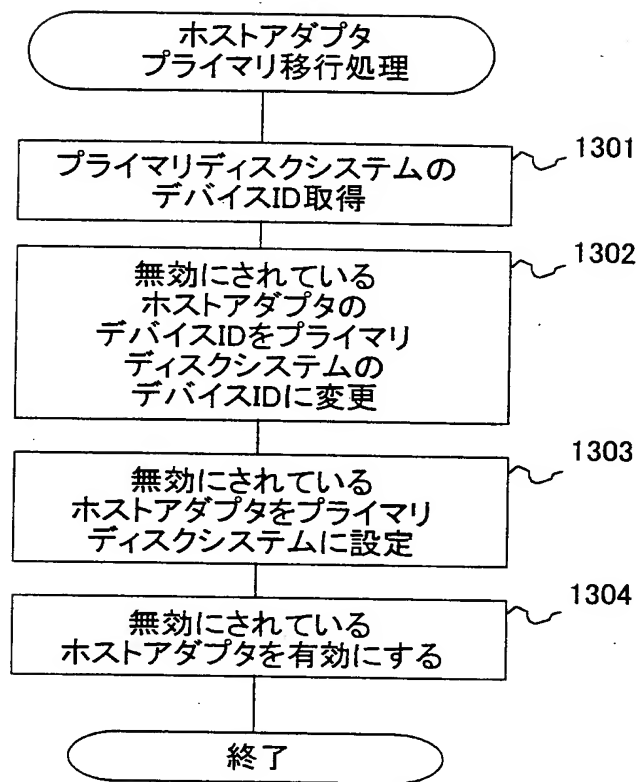


図14

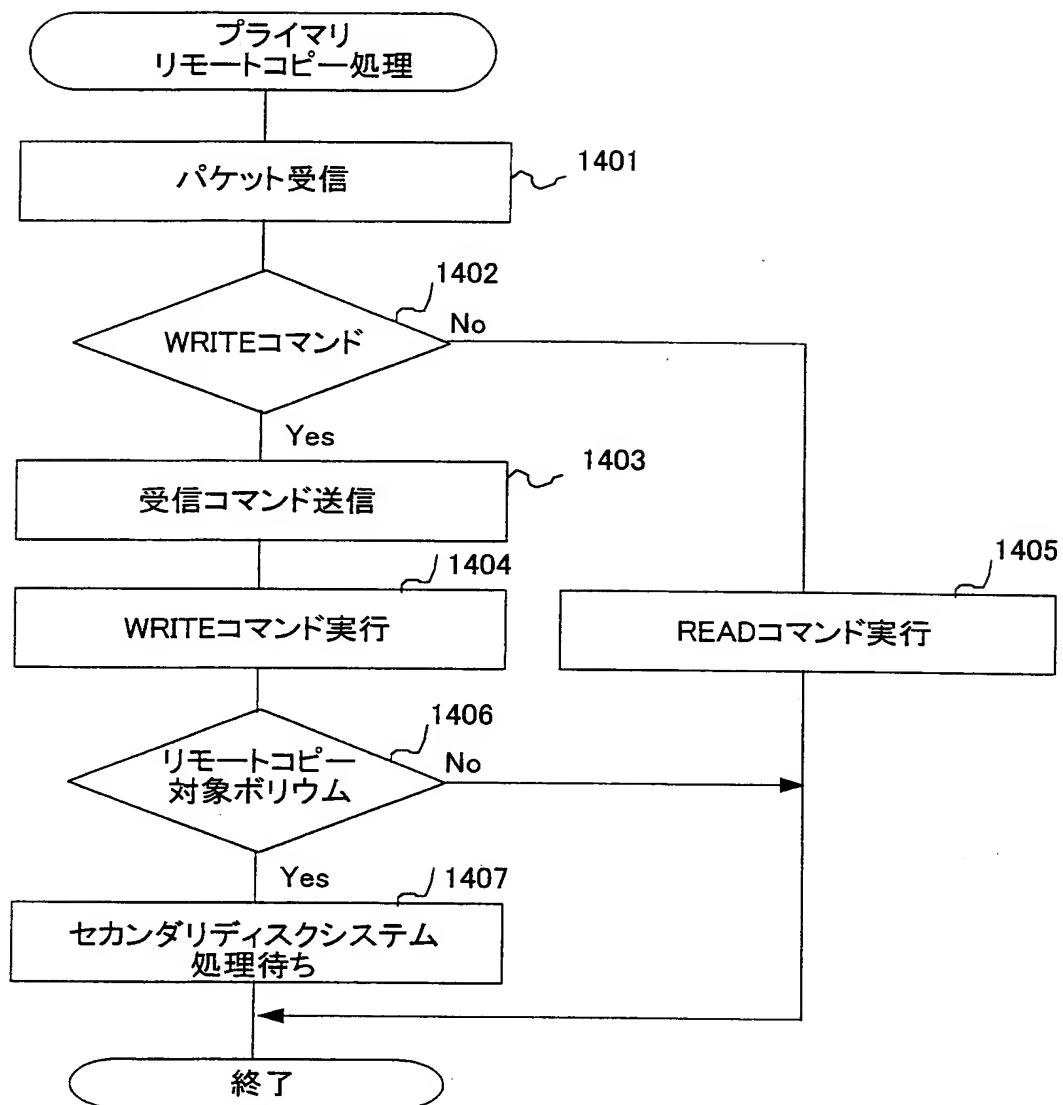


図15

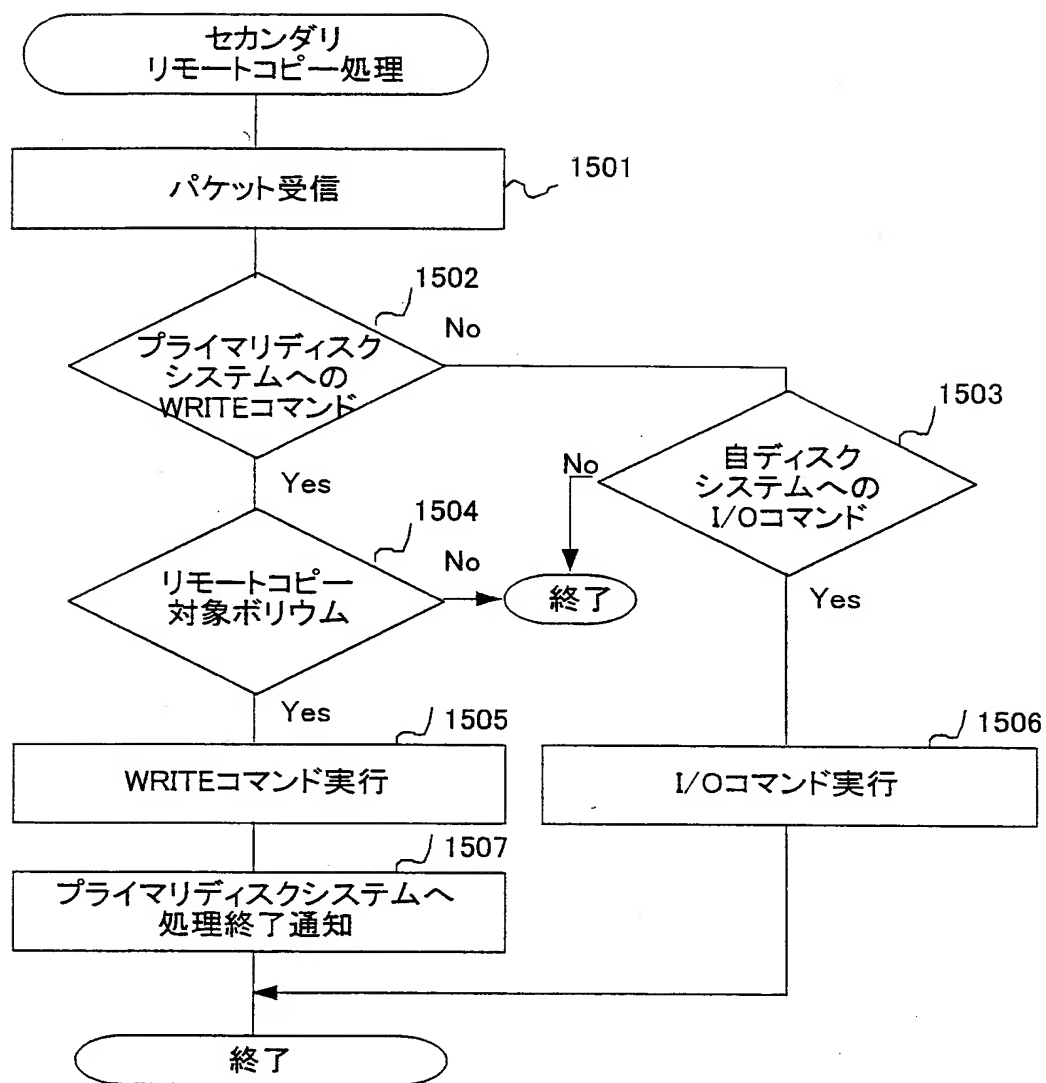


図16

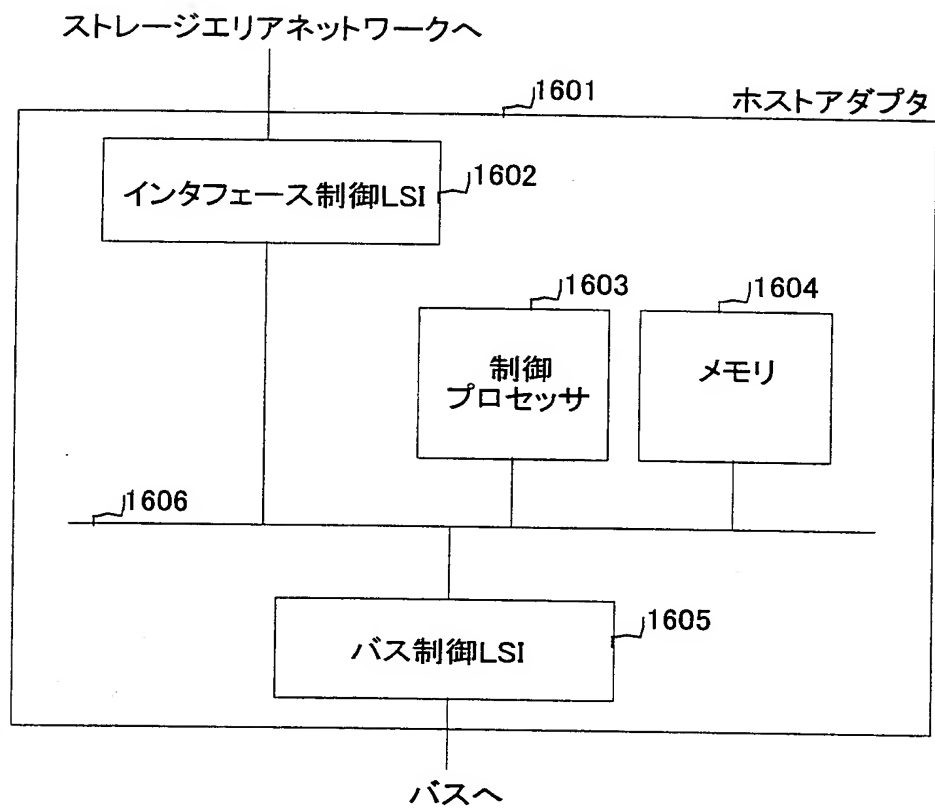


図17

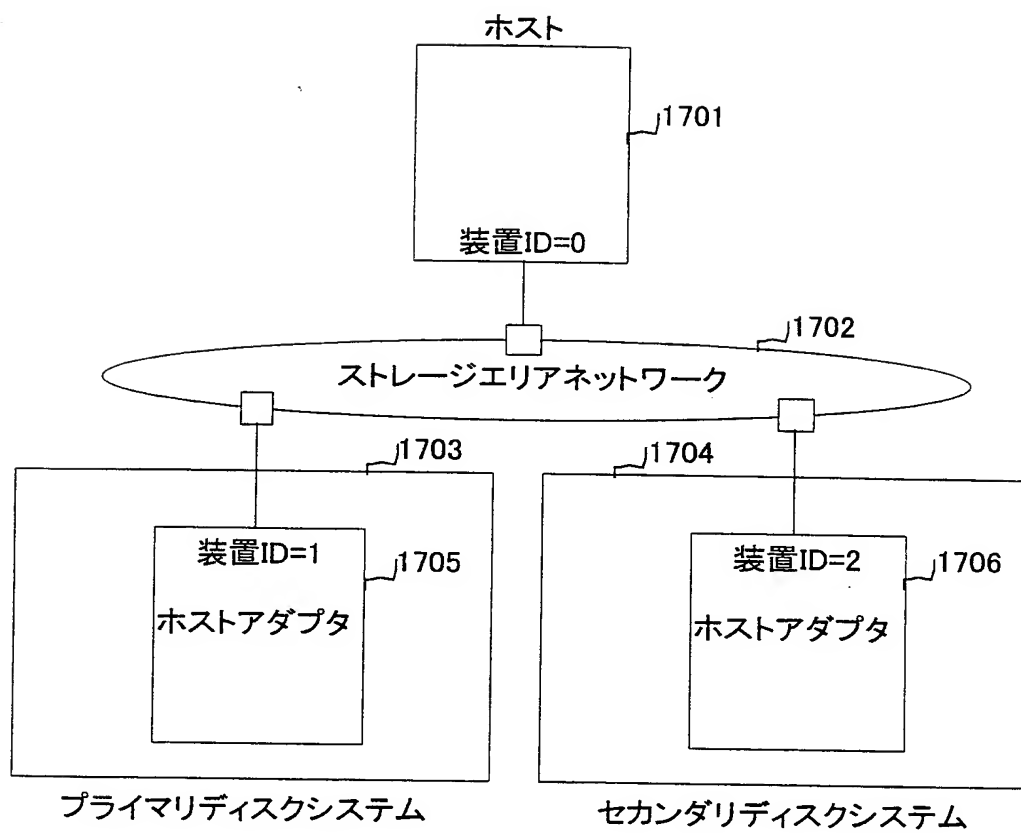


図18

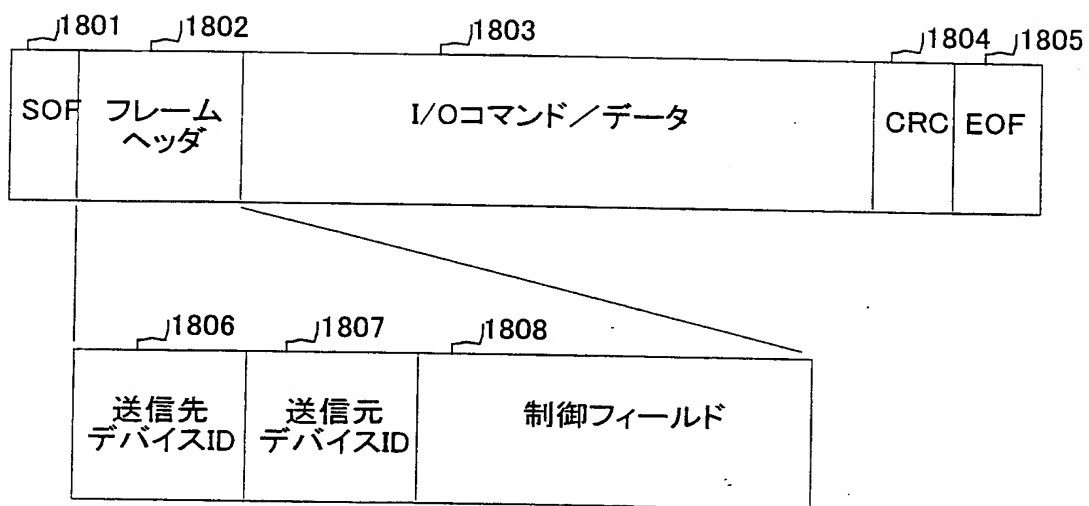


図19

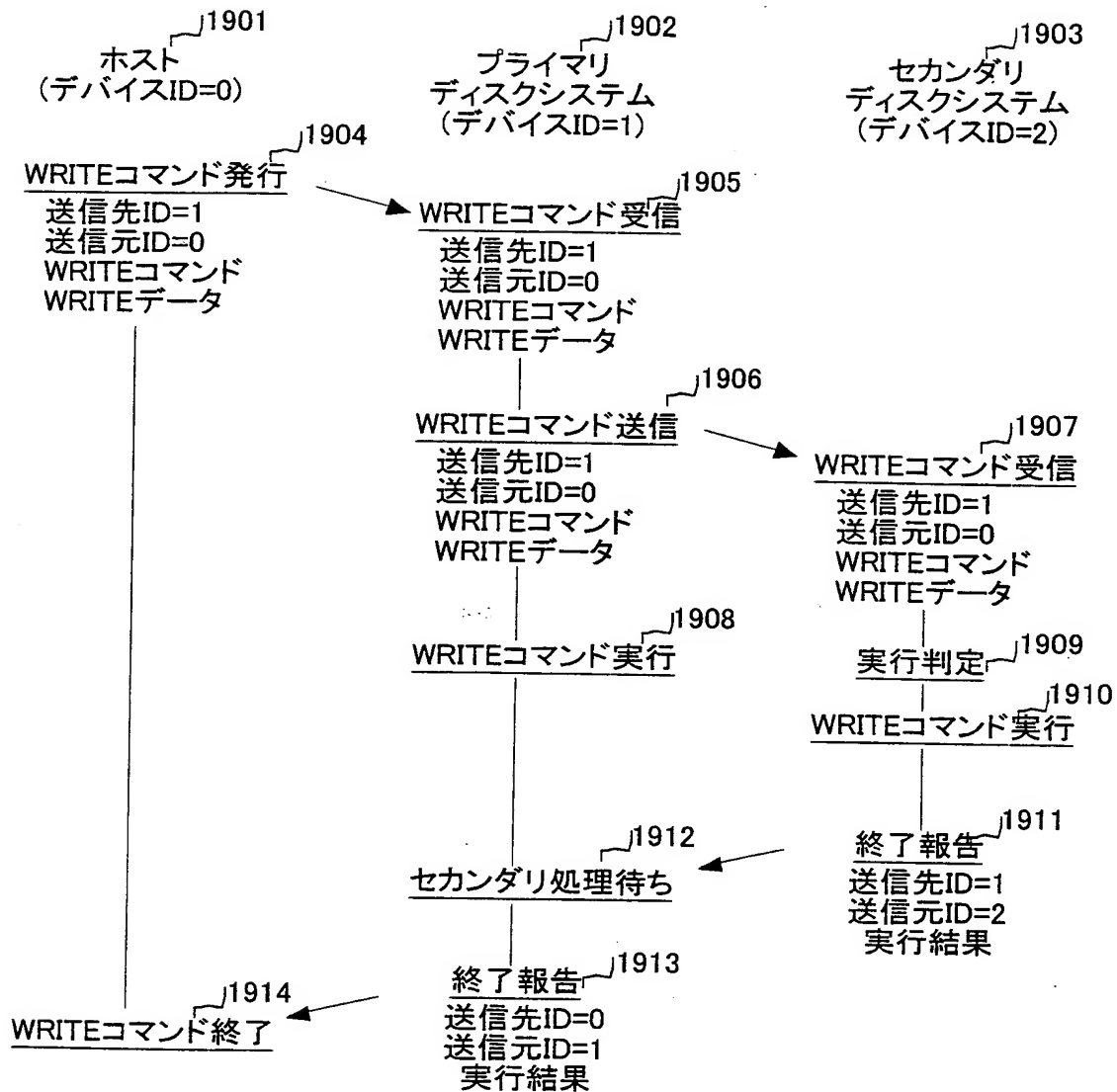
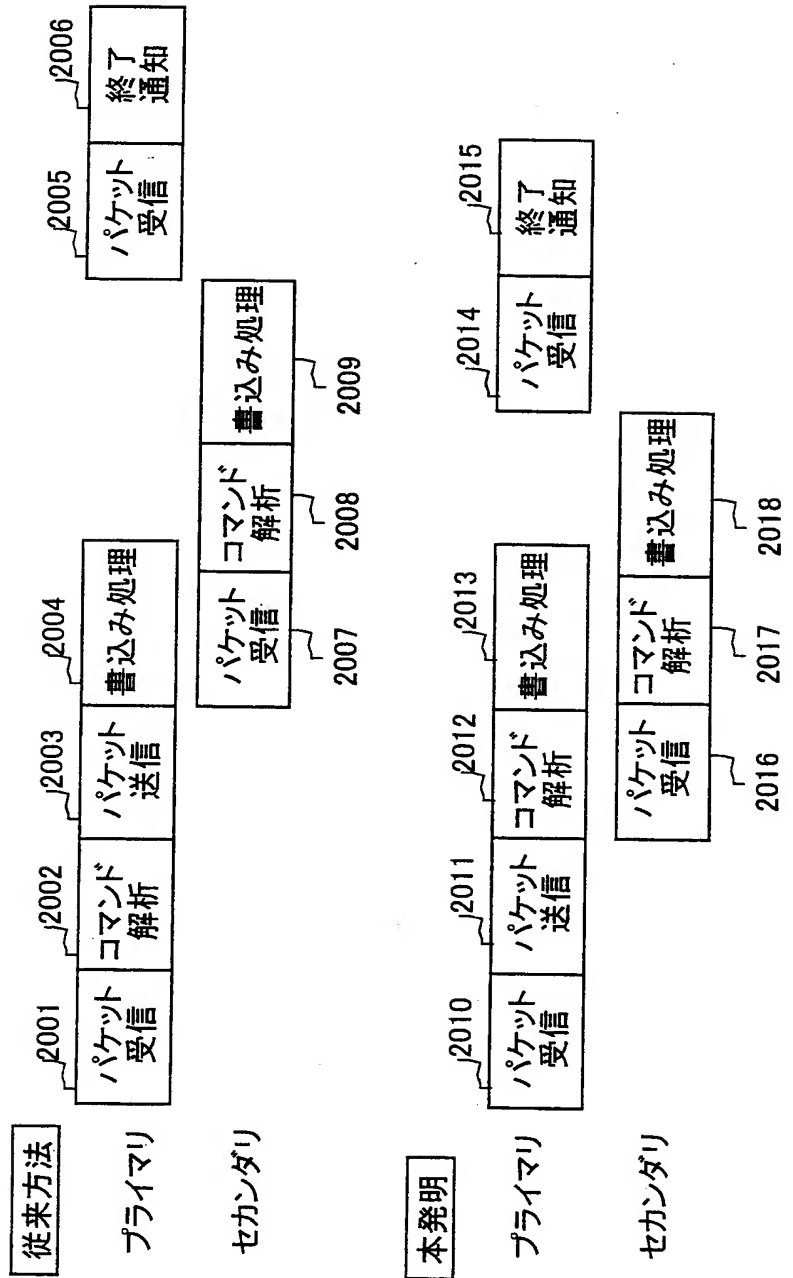


図20



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP99/05850

A. CLASSIFICATION OF SUBJECT MATTER Int.Cl ⁷ G06F3/06 Int.Cl ⁷ H04L29/14 Int.Cl ⁷ G11B20/10 Int.Cl ⁷ G06F12/00 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) Int.Cl ⁷ G06F3/06 Int.Cl ⁷ H04L29/14 Int.Cl ⁷ G11B20/10 Int.Cl ⁷ G06F12/00 Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Jitsuyo Shinan Koho 1926-1996 Jitsuyo Shinan Toroku Koho 1996-2000 Kokai Jitsuyo Shinan Koho 1971-2000 Toroku Jitsuyo Shinan Koho 1994-2000 Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP, 10-93556, A (Mitsubishi Electric Corporation), 10 April, 1998 (10.04.98) (Family: none)	1-8
A	US, 5901280, A (Hitachi, Ltd.), 03 March, 1995 (03.03.95), & JP, 7-56842	1-8
A	JP, 9-146812, A (Sanyo Electric Co., Ltd.), 06 June, 1997 (06.06.97) (Family: none)	4
A	JP, 5-265829, A (Hitachi, Ltd.), 15 October, 1993 (15.10.93) (Family: none)	7-8
A	JP, 5-233514, A (NEC Eng. Ltd.), 10 September, 1993 (10.09.93) (Family: none)	7-8
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier document but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family	
Date of the actual completion of the international search 18 January, 2000 (18.01.00)	Date of mailing of the international search report 01 February, 2000 (01.02.00)	
Name and mailing address of the ISA/ Japanese Patent Office	Authorized officer	
Facsimile No.	Telephone No.	

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int, Cl⁷ G06F3/06 Int, Cl⁷ H04L29/14
 Int, Cl⁷ G11B20/10
 Int, Cl⁷ G06F12/00

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int, Cl⁷ G06F3/06 Int, Cl⁷ H04L29/14
 Int, Cl⁷ G11B20/10
 Int, Cl⁷ G06F12/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報 1926-1996年
 日本国公開実用新案公報 1971-2000年
 日本国実用新案登録公報 1996-2000年
 日本国登録実用新案公報 1994-2000年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	JP, 10-93556, A(三菱電機株式会社), 10, 4月, 1998(10. 04. 98), (ファミリーなし)	1-8
A	US, 5901280, A(株式会社日立製作所), 3, 3月, 1995(03. 03. 95)&JP, 7-56842	1-8
A	JP, 9-146812, A(三洋電機株式会社), 6, 6月, 1997(06. 06. 97), (ファミリーなし)	4
A	JP, 5-265829, A(株式会社日立製作所), 15, 10月, 1993(15. 10. 93), (ファミリーなし)	7-8
A	JP, 5-233514, A(日本電気エンジニアリング株式会社), 10, 9月, 1993(10. 09. 93), (ファミリーなし)	7-8

☐ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」 特に関連のある文献ではなく、一般的技術水準を示すもの

「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)

「O」 口頭による開示、使用、展示等に言及する文献

「P」 国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」 同一パテントファミリー文献

国際調査を完了した日

18. 01. 00

国際調査報告の発送日

01.02.00

国際調査機関の名称及びあて先

日本国特許庁 (ISA/J P)

郵便番号 100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

三好洋治



5E

9564

電話番号 03-3581-1101 内線 3520